A TREE-STRUCTURED SURVIVAL MODEL WITH INCOMPLETE AND TIME-DEPENDENT COVARIATES: ILLUSTRATIONS USING TYPE 1 DIABETES DATA

by

Shui Yu

BSc, Central China Normal University, 1991

MSc, Northeastern Normal University,1994

MS, University of Pittsburgh, 2002

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Shui Yu

It was defended on

November 15, 2006

and approved by

Dissertation Advisor: Sati Mazumdar, Ph.D.
Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member: Dorothy Becker M.B.B.Ch.
Professor
Director, Division of Endocrinology and Diabetes
Children's Hospital and University of Pittsburgh

Committee Member: Nancy B. Sussman, Ph.D.
Assistant Professor
Department of Environmental and Occupational Health
Graduate School of Public Health
University of Pittsburgh

Committee Member: Howard E. Rockette, Ph.D.
Professor and Chair
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member: Ingrid Libman, M.D, Ph.D.
Assistant Professor
Division of Pediatric Endocrinology and Diabetes
Children's Hospital and University of Pittsburgh

Sati Mazumdar, Ph.D.

A TREE-STRUCTURED SURVIVAL MODEL WITH INCOMPLETE AND TIME-DEPENDENT COVARIATES: ILLUSTRATIONS USING TYPE 1 DIABETES DATA

Shui Yu, PhD

University of Pittsburgh, 2006

A tree-structured recursive partitioning algorithm is adapted for censored survival analysis with incomplete and time-dependent covariates. The only assumptions required for this method are those that guarantee identifiability of the conditional distribution of the survival time given the covariates, providing broad applicability. The method also provides personalized prognosis. A conditional incremental imputation procedure, which does not depend on any model assumptions, is implemented to impute missing covariate values. These novel algorithms are applied to assess the role of islet antibodies (ICAs) as predictive markers for Type 1 diabetes mellitus (T1DM) progression in a longitudinal study of 300 first-degree relatives (FDRs) that were consecutively enrolled between 1977 through 2001 from the Children's Hospital of Pittsburgh Registry. Results provide evidence that ICAs predict a more rapid progression to insulin-requiring diabetes in GAD65 positive relatives. A cross-validation study confirms the findings. Islet-cell antibodies (ICAs) are important markers of Type 1 diabetes. The issue regarding whether or not the measurement of ICAs should be completely replaced by biochemical markers detecting islet autoantibodies (AAs) for the prediction of T1DM has been the subject of endless debates. Our conclusion that ICAs should remain part of the assessment of T1DM risk is of great public health significance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to give special thanks to Dr. Sati Mazumdar, Chair of my committee, and Dr. Dorothy Becker, my Supervisor at the Children's Hospital of Pittsburgh, for mentoring me through the long dissertation process with vision, patience, understanding, and encouragement. I would also like to thank my committee members, Dr. Ingrid Libman, Dr. Howard E. Rockette, and Dr. Nancy B. Sussman, for their unwavering support, guidance, and invaluable input through this challenging undertaking.

The Biostatistics Department of the University of Pittsburgh as a whole provided an excellent resource in various ways and I feel extremely lucky to be able to be associated with such an outstanding group of people.

Special thanks to Max and Susan Pietropaolo, Polly Swanson, Karen Riley, and all the other colleagues of mine at the Children's Hospital of Pittsburgh for all their help and friendship. I would also like to thank Nandita Mukhopadhyay from Human genetics Department of the University of Pittsburgh for helping me with the programming. Dr. Bacchetti and Dr. Segal kindly shared their program.

Finally, I extend heartfelt thanks to everyone who helped me to get where I am today, I can not make it without you.

1.0 INTRODUCTION

Clinical prognostic models are complex tools that combine patient characteristics to predict clinical outcomes. They are very important in difficult clinical decision makings, such as selecting patients for intervention therapy and optimal timing for such interventions. In medical decision makings, identification of groups of patients with differing time to a selected event (such as conversion to a disease or death) is often desired to understand the relationship between patient characteristics and survival. The objective of a survival study is to identify the relationship between treatments, risk factors and the time to event. Hence survival analysis is commonly used for this purpose.

In survival analysis, Cox's proportional hazards model (Cox, 1972) is widely used. Although it is a flexible tool for the study of covariate associations with survival time, it does not directly lead to models for prognostic groups. Interactions in the Cox model are often modeled artificially and may not reflect meaningful clinical situations. Besides, the Cox model is sometimes used arbitrarily without proper model validation checks. Over the years, many authors have noted violations of the proportional hazards assumption in various applications (Lancaster and Nickell, 1980; Gail, Wieand, and Piantadosi, 1984; Struthers and Kalbfleisch, 1986; Ford, Norrie, and Ahmadi, 1995). This occurs often under clinical settings when important prognostic variables are used to predict survival of the patients.

The goal of tree-structured methods is the identification of meaningful subgroups that are expressed as logical combinations of covariate values. This is appealing in biomedical settings since it translates into finding groups of patients with similar prognoses, which can be used for building prognostic models. Assignment/classification of new patients to different prognostic groups becomes effortless and involves just answering a sequence of yes/no questions. Such groupings, characterized by common risk factor values, are important in making treatment decisions, assessing disease heterogeneity and for subsequent covariate adjustment so as to facilitate treatment comparisons. The tree diagram can also display statistical summaries of the groups permitting easy prediction for a specific patient.

In many instances, time-dependent covariates are used in survival analysis. For example, blood pressure, weight, disease history, and blood antibody levels may be collected at selected periodic time points, and treatment or other factors may change over time. The use of time-dependent covariates offers exciting opportunities for exploring associations and potentially causal mechanisms that may lead to dynamic prognosis in which the relative risk can change from one time point to the next as the values of the covariates change.

The purpose of this dissertation is to construct a tree-structured time-dependent prognostic model, and use this model to analyze a dataset for Type 1 diabetes based on longitudinal follow-up from a large group of first-degree relatives of Type 1 diabetes patients, and subsequently to discuss the reliability of this model. The proposed method will provide an exploratory tool for the analysis of large survival datasets with complex data structure which involves follow-up and serial measurements of patient characteristics over time.

2.0 LITERATURE REVIEW

One of the areas of great methodological advances in biostatistics has been the ability to handle censored time-to-event data. "Censored" means that some units of observation are observed for some lengths of time but do not experience the event (or endpoint) under study. Kaplan and Meier presented the product limit or Kaplan-Meier method to efficiently use all of the data to estimate the time-to-event curve (Kaplan and Meier, 1958). Comparison of groups based on this nonparametric estimate is given by the log rank test. Cox proposed a model which puts predictor/explanatory variables into consideration (Cox, 1972). This model is based on the hazard function which may be thought of as the instantaneous probability of an event at a particular time. The effects of the covariates are estimated by multiplying the hazard function by a function of the explanatory covariates. This means that two units of observation have a ratio of their hazards that is constant over time and depends on their covariate values. This model is usually called the Cox regression model or the proportional hazards regression model. It is often used to examine the predictive value of survival in terms of subject (often patients in medical setting) covariates such as treatment, age, gender, height, weight, relative weight, smoking status, ethnicity categories, diastolic or systolic blood pressures, education, and income, to predict survival. The exponential of the coefficients from the Cox model gives the relative risk for an increase of one unit for the covariate in question. At this time, the Cox model is probably the most widely used model in this area.

Recently, there has been growing interest in using tree-structured models in survival analysis for both statistical and clinical reasons. Statistically, they are applicable to more general situations than classical regression approaches; clinically, they meet the demand of the investigators who are usually interested in grouping patients with differing prognoses. A tree-based method provides a clear description of complex interactions amongst prognostic factors and does not depend on common but often unrealistic assumptions, such as linearity of effects for continuous variables. Tree-structured models also identify effects of covariates inside a subgroup while the conventional models characterize covariate effects across the entire sample.

Original tree-structured models were used in classification and regression situations by Morgan and his colleagues (Morgan and Sonquist, 1963). Advances in the practical and theoretical aspects of tree-based methods were developed by Breiman and his colleagues in their monograph Classification and Regression Trees (CART) (Breiman et al, 1984). Generally speaking, tree-based methods recursively partition the covariate space into disjoint regions and assign the corresponding data into groups (nodes). For each node to be split, some measure of separation in the response distribution between the two daughter nodes is calculated. All possible splits for each of the covariates are evaluated, and the variable to be split and the split point are chosen that best separate the nodes. The same procedure is applied recursively to increase the number of nodes until each contains only a few subjects. The resulting model is represented as a binary tree.

Gordon and Olshen (1985) presented the first adaptation of CART to censored survival data, using distance measures between nearest continuous approximation of Kaplan-Meier curves. Davis and Anderson (1989) proposed a method based on the exponential log likelihood at nodes. Therneau, Grambsch, and Fleming (1990) proposed a method in which Martingale

residuals were used directly in the CART regression algorithm with squared error loss. LeBlanc and Crowley (1992) extended the proportional hazards regression to tree-structured relative estimates for censored survival data with a one-step full likelihood estimation procedure. All four of these methods are based on measuring homogeneity within a node so that the application of CART is straightforward. Segal (1988) presented a totally nonparametric application using the Harrington-Fleming (1982) classes of two-sample rank statistics that based the partitioning on between-node separation instead of within-node homogeneity. Later, Bacchtti and Segal (1995) further extended this method to allow for truncation and time-dependent covariates. LeBlanc and Crowley (1993) developed a recursive partitioning procedure based on maximizing the dissimilarity in the survival distributions of patients between regions of the covariate space.

These last two methods are based on maximizing the dissimilarity in survival distributions between different regions of the covariate space. Ahn and Loh (1994) developed a tree-structured proportional hazards regression model that stratifies data according to selected covariate values and fits separate proportional hazards models to each stratum. Most of the exiting survival trees methods are suitable for only time-independent covariates. In this dissertation, we extend tree-structured modeling further to accommodate survival data with incomplete and time-dependent covariates. A nonparametric method of imputing missing covariate values was also implemented. So we developed a more general approach to analyze data even when little information about the distributions of survival time and/or covariates is available.

3.0 GENERAL TREE-STRUCTURES REGRESSION METHOD

Tree-structured regression methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures. The applications of these methods in medicine are far reaching. They are considered to represent complex diagnostic and treatment strategies in a way that is ideally suited to mimic actual thinking processes. This section presents a simplified description of regression trees to facilitate understanding of the subsequent extensions. For a more detailed understanding, the CART monograph is recommended (Breiman et al., 1984). In this section, attention is restricted to the regression setting. The data structure for this setting can be defined as the following: suppose that we have observed p covariates, denoted by a p-vector $\mathbf{x}$, and a response $\mathbf{y}$ for n individuals, then for the i_{th} individual, the measurements are:

$\mathbf{x}_i = (x_{i1}, \cdots x_{ip})'$ and y_i, $i = 1, \cdots, n$. Here y can be either continuous or discrete, but uncensored. The objective is to model the probability distribution $P(\boldsymbol{y} \mid \mathbf{x})$.

In order to construct a regression tree based on the data, four components are required. These are:

1. **A set of (binary) questions of the form "Is $\mathbf{x} \in$ A?" where x is a case and $\mathbf{A} \subset \chi$, the predictor space.** The answer to such a question induces a partition, or split on the predictor space. Cases for which the answer is "yes" are associated with the region A and those for which the answer is "no" are associated with the complement of A. The subsamples so formed are called nodes.

2. **A goodness-of-split criterion ϕ(s,t) that can be evaluated for any split s of any node.**
   The criterion is used to assess the worth of the competing splits.
3. **Some means to determine the appropriate tree size.** It can be done either by stopping the splitting process when some criteria is met or by growing a large tree and use some criteria to prune it.
4. **Statistical summaries of the terminal nodes.**
   The above splitting process is recursively repeated for the resulting subgroups until it is decided that no further split is needed.

4.0 A TREE-STRUCTURED MODEL FOR SURVIVAL DATA WITH INCOMPLETE AND TIME-DEPENDENT COVARIATES

In this section, we outline the extension of the tree-structured regression method to accommodate right-censored survival data, incomplete time-dependent covariates.

4.1 EXTENSIONS TO RIGHT CENSORED DATA

The survival data can be represented by $\boldsymbol{T}$ denoting the time to an event, which can be death or the occurrence of a disease. For a variety of reasons including lost to follow-up and the limited period of a study, we may not be able to observe $\boldsymbol{T}$ until the event occurs for everyone in the study. Thus, what we actually observe is a censored time $\boldsymbol{C}$ which is smaller than or equal to $\boldsymbol{T}$. Let $\boldsymbol{Y} = min\ (\boldsymbol{T, C})$. The question is how to facilitate the censored time $\boldsymbol{Y}$ in the tree-structured methods. This is done by modifying the splitting and pruning criteria (Bacchetti and Segal, 1995; LeBlanc and Crowley, 1993)

Instead of using goodness-of-split criteria for uncensored data, censored data rank statistics are used. These statistics can be calculated as the following:

For the i_{th} event time $y_{i,}$ we define:

n_i = # of subjects with event time $y_{i,}$;

n_{i1} = # of subjects in the left node;

m_{i1} = # of subjects with event;

a_i = # of events in the left node;

Similar numbers can be calculated for each of the distinct event times and the statistics takes the following form:

$$TW = \frac{\sum_{i=1}^{k} w_i[a_i - \mathrm{E}_0(A_i)]}{[\sum_{i=1}^{k} w_i^2 \mathrm{var}_0(A_i)]^{1/2}},$$

where A_i is the random variable corresponding to number of events in the left node for the i_{th} table; w_i are constants used to weight the respective tables; the sum is over all tables, i.e., all distinct observations; the null hypothesis is that the event rates for the two nodes are equal. For fixed margins the null expectations and variances are hypergeometric:

$$\mathrm{E}_0(A_i) = \frac{m_{i1} n_{i1}}{n_i}$$

$$\mathrm{var}_0(A_i) = \left[\frac{m_{i1}(n_i - m_{i1})}{n_i - 1}\right]\left[\left(\frac{n_{i1}}{n_i}\right)\left(1 - \frac{n_{i1}}{n_i}\right)\right]$$

Different statistics can be obtained by setting weight W_i to different numbers:
1. $W_i = 1$ gives log-rank statistic (Peto and Peto, 1972).
2. $W_i = n_i$ gives Gehan statistic (Gehan, 1965).

4.2 EXTENSIONS TO TIME-DEPENDENT COVARIATES

Bacchetti and Segal extended the tree- building procedure to allow for a time-dependent covariate X_j (t) (Bacchetti and Segal, 1995). First, a split based on a question of the form "Is $X_{jk}(t) \leq c$? " for some specific value c is considered. Subjects k with $x_{jk}(t) > c$ at all times go to the right node, subjects with $x_{jk}(t) \leq c$ at all times go to the left node, but subjects with $x_{jk}(t) \leq c$ for some time and $x_{jk}(t) > c$ at other time need to contribute to the left node some of the time and to the right node at other times. They first considered the case when X_j (t) is nondecreasing in t and t_k^* is the last time when $x_{jk}(t) \leq c$, with $r_k < t_k^* < y_k$.

Proper testing of the split requires that subject k to be considered part of left node at failure times such that $r_k < t_i \leq t_k^*$ and part of right node when $t_k^* < t_i \leq y_k$. Subject k's survival experience can be regarded as being composed of the non-overlapping survival experience of two pseudo-subjects k_1 and k_2 . Pseudo-subjects k_1 is only at risk up to time t_k^* , i.e., is right-censored at t_k^* , while k_2 is not at risk until after t_k^* , i.e., is left-truncated at t_k^* . The split is handled analogously when X_j (t) is nonincreasing.

General time-dependent covariates can be accommodated by splitting observations into more than two psedudo-observations. The process is summarized in Figure 4.1.

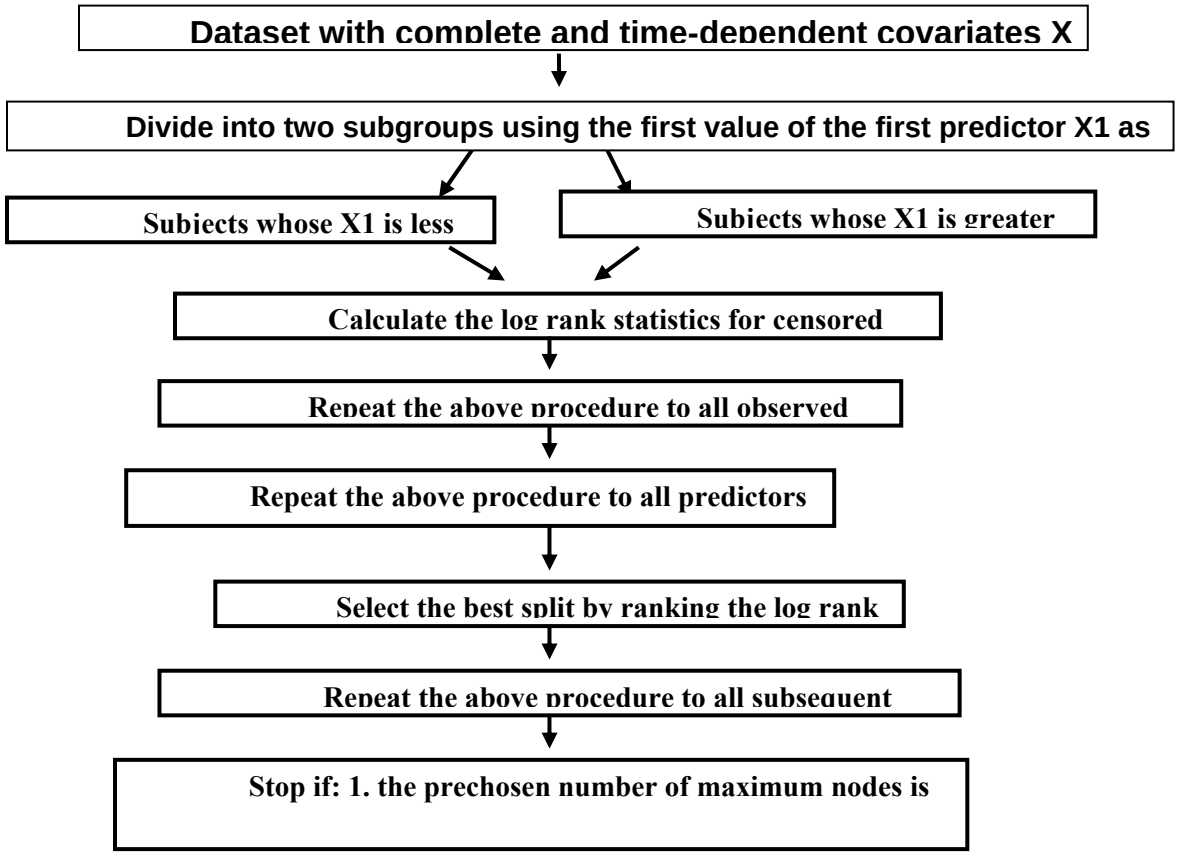

Figure 4-1 Flow chart of the tree building process

4.3 IMPUTATION OF MISSING COVARIATE VALUES

In clinical research, often some covariate values are found to be missing for some patients for deferent reasons. For example, a patient may refuse to answer a question, a biochemical analysis fails or a case report form is lost. The standard approach incorporated in most statistical software packages is the complete case analysis, that is, all subjects with at least one missing

value in the covariates are omitted from the analysis. This approach is of course wasteful of information, as the omitted subjects carry information on the relation between the observed covariates and the outcome variable. This practice leads to a reduction in the statistical power, and may also lead to biased results. There exist some simple methods to use all subjects, for example, by imputing means for the missing values from values of similar subjects or by regarding missing values as an additional category. However, these approaches can be a source of serious bias or can result in an overestimation of the gained precision. Hence a recommendation for such methods is only possible in special situations, not in general.

Missing values can be imputed in cases where the reason for the data being missing is known, or when it can be explained by the available data. This will increase the power of an analysis and may produce models that are statistically more reliable and applicable within clinical practice. To establish reliably the effects of different prognostic factors on long-term survival, in this section, we describe a method to deal with the missing data. An iterative use of tree based models for missing data imputation procedure is used for this purpose. This procedure was proposed by (Conversano et al., 2004). We chose this method for the following reasons:

a) Its nonparametric nature;
b) Its flexibility, because it handles simultaneously categorical and numerical predictors and interactions among them;
c) Its simplicity.

4.3.1 Basic Idea

Given a variable for which data are missing, and set of other d ($d < p$)variables are observed, the method works by using the former as the response variable $\boldsymbol{y}$ and the latter as covariates $\mathbf{x}_1, \mathbf{x}_{2,\ldots,}$

$\mathbf{x}_d$. The resulting tree model explains the distribution of the response variable in terms of the values of the covariates. Since the terminal nodes of the tree are homogeneous with respect to the $\mathbf{x}$'s, they provide candidate imputation values. To deal with the data presenting missing values in many covariates, an incremental approach based on a suitably defined data preprocessing schema is used.

4.3.2 Implementation

Let $\boldsymbol{X}$ be the original n×p data matrix, with d completely observed variables, and q covariates with missing data. We perform a two-way rearrangement of $\boldsymbol{X}$, one with respect to the columns (X_1, $X_{2,\dots}$ $X_{d,\dots}$ X_p)and one with respect to the rows (1,2,…m,…,n) using a lexicographic ordering that matches the ordering by value, corresponding to the number of missing values occurring in each record. Practically, we form a string vector of length n that indicates the occurrence and the number of missing values for each row of $\boldsymbol{X}$. This allows to order $\boldsymbol{X}$ in a way that the first incomplete column X_d presents the lowest number of missing values and it follows the complete observed ones. Furthermore, columns also are ordered in the way that the first m rows contain instances with no missing values and the remaining (n-m) rows present missing values. As a result, $\boldsymbol{X}$ is partitioned into four disjoint matrices as follows:

$$X_{n,p} = \begin{bmatrix} A_{m,d} & C_{m,p-d} \\ B_{n-m,d} & D_{n-m,p-d} \end{bmatrix}$$

Note that, as a consequence of the ordering schema, only $\mathbf{D}$ contains missing values while the other three blocks are completely observed with respect to their rows and columns.

4.3.3 Incremental Imputation

The missing data imputation is iteratively done using tree-structured models. With respect to the records presenting only one missing value, a simple tree is used. Here, the variable with missing values is the response and the other observed variables are the covariates. The tree is built on the current complete data cases in $\mathbf{A}$ and its results are used to impute the cases in $\mathbf{D}$. In fact, terminal nodes of the tree represent candidate "imputed values". Actual imputed values are obtained by dropping down the tree the cases of $\mathbf{B}$ corresponding to the missing values in $\mathbf{D}$ (for the variable under imputation), till a terminal node is reached. The conjunction of the filled-in cells of $\mathbf{D}$ with the corresponding observed rows in $\mathbf{B}$ generates new records which are appended to $\mathbf{A}$, that gains the rows whose missing values have been just imputed and a "new" column corresponding to the variable under imputation.

For records presenting multiple missing values, trees are used iteratively. In this case, according to the previously defined lexicographic ordering, the tree is first used to fill in the missing values of the covariate presenting the smallest number of incomplete records. The procedure is then repeated for the remaining covariates under imputation. In this way, we form as many trees as the number of covariates with missing values. This algorithm is presented in Figure 4.2.

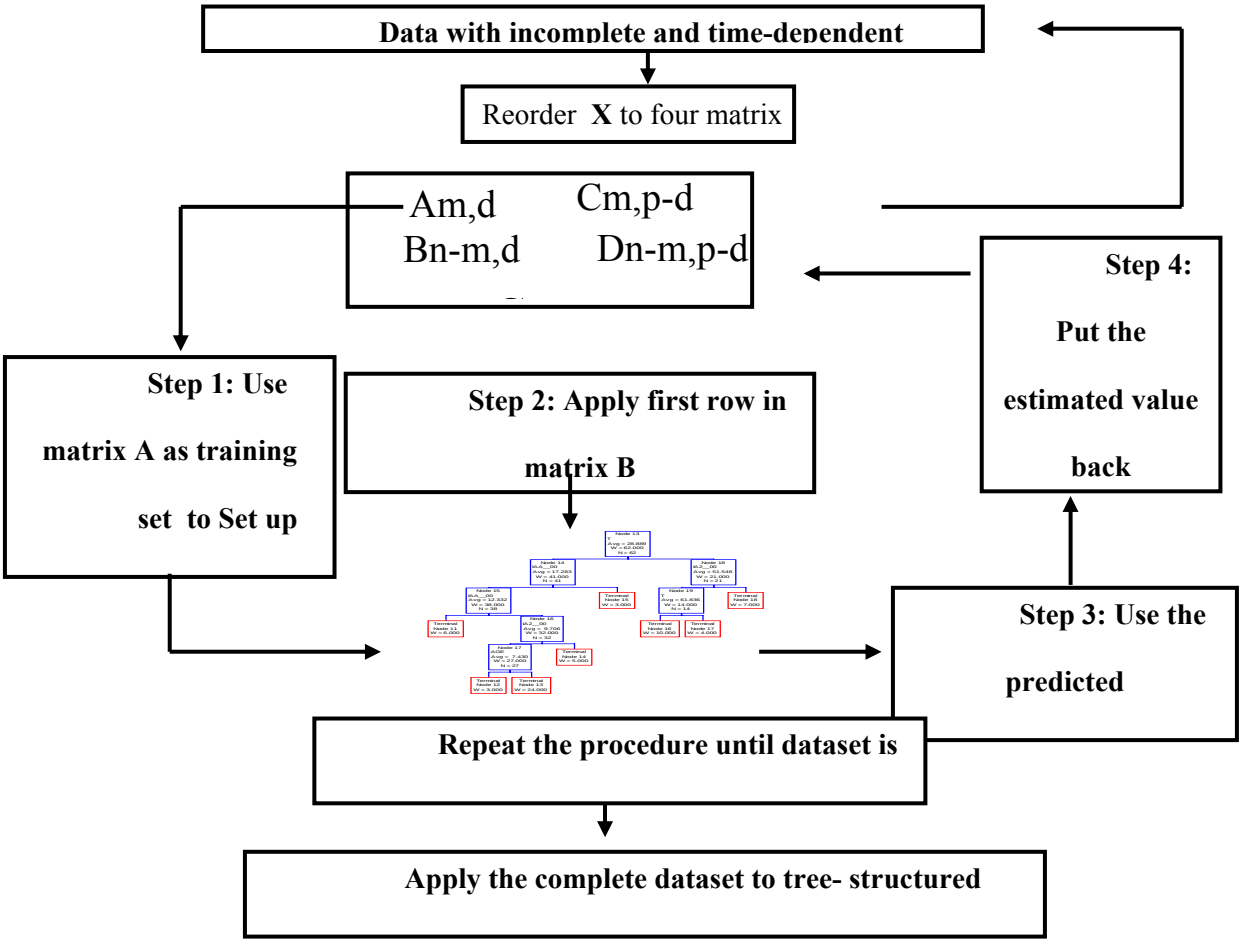

Figure 4-2 Flow chart of the missing covariate imputation algorithm

4.4 AN EXAMPLE OF THE IMPUTATION ALGORITHM

A part of the real dataset is used to demonstrate how to use the algorithm and CART to estimate the missing covariates. Figure 4.3 is an example of the original dataset with some missing covariates. Figure 4.4 shows the reorganized and partitioned dataset .

PID	age	race	gender	RAT1	RAT2	RAT3	GAD	IA2
19,340,101	27	0	1	80	220	134	1.023	-0.009
21,840,101	29	0	1	160	160	320	-0.003	-0.007
28,520,101	43	0	1	270	160	160	0.008	-0.004
27,690,101	41	0	1	270	160	#MISSING	0.816	-0.002
25,960,101	27	0	1	80	240	#MISSING	0.007	0.001
16,020,101	37	0	1	10	160	#MISSING	0.027	0.001
440,105	24	0	1	80	134	160	0.047	0.005
83,700,106	27	0	0	80	#MISSING	320	0.020	0.006
20,570,101	30	1	1	80	#MISSING	160	1.041	0.006
19,830,104	10	0	0	320	#MISSING	106	0.935	0.008
740,101	40	0	1	0	160	160	0.917	0.008
6,200,106	12	0	1	320	320	106	1.355	0.015
3,070,101	47	0	1	10	160	#MISSING	0.005	0.026
250,101	35	0	1	320	320	#MISSING	1.182	0.457
6,100,101	36	0	1	80	160	#MISSING	0.382	0.543
23,330,103	18	0	0	80	106	#MISSING	0.032	0.863
5,190,103	10	0	0	80	160	270	0.452	1.038
18,670,104	15	0	1	180	106	270	0.642	1.070
53,740,103	36	0	1	320	320	80	1.665	1.142
16,970,102	31	1	0	80	220	10	0.676	1.145

Figure 4-3 An example dataset for missing data imputation algorithm

PID	age	race	gender	GAD	IA2	RAT1	RAT2	RAT3
6,200,106	12	0	1	1.355	0.015	320	320	106
250,101	35	0	1	1.182	0.457	320	320	#MISSING
53,740,103	36	0	1	1.665	1.142	320	320	80
25,960,101	27	0	1	0.007	0.001	80	240	#MISSING
19,340,101	27	0	1	1.023	-0.009	80	220	134
16,970,102	31	1	0	0.676	1.145	80	220	10
21,840,101	29	0	1	-0.003	-0.007	160	160	320
28,520,101	43	0	1	0.008	-0.004	270	160	160
27,690,101	41	0	1	0.816	-0.002	270	160	#MISSING
16,020,101	37	0	1	0.027	0.001	10	160	#MISSING
740,101	40	0	1	0.917	0.008	0	160	160
3,070,101	47	0	1	0.005	0.026	10	160	#MISSING
6,100,101	36	0	1	0.382	0.543	80	160	#MISSING
5,190,103	10	0	0	0.452	1.038	80	160	270
440,105	24	0	1	0.047	0.005	80	134	160
23,330,103	18	0	0	0.032	0.863	80	106	#MISSING
18,670,104	15	0	1	0.642	1.070	180	106	270
83,700,106	27	0	0	0.020	0.006	80	#MISSING	320
20,570,101	30	1	1	1.041	0.006	80	#MISSING	160
19,830,104	10	0	0	0.935	0.008	320	#MISSING	106

Figure 4-4 Reorganized and partitioned example dataset

5.0 A REAL-LIFE EXAMPLE: TYPE1 DIABETES STUDY

Diabetes is a rapidly growing health problem. Currently, more than 18 million people in the United States have diabetes. Type 1 diabetes is a life-long disorder that can arise in children or adults. People with Type 1 diabetes need insulin replacement for the rest of their lives. Most patients need multiple daily injections or an insulin pump for good control of their blood glucose. If Type 1 diabetes could be prevented or delayed, millions of people across the globe would enjoy longer lives, improved health, and freedom from the burden of managing this difficult disease.

In the past 20 years, researchers have learned a great deal about the factors that contribute to diabetes risk. Table 5.1 and Figure 5.1 show some of the factors that contribute to Type 1 diabetes risk and their sensitivity and positive predictive values associated with them. With this information, they can now be used to identify people at risk for Type 1 diabetes and to design clinical trials to test strategies to prevent or delay onset of disease. Table 5.1 presents some of the assays for the auto antibodies associated with Type 1 diabetes.

Table 5-1 Islet Cell Autoantibody Assays*

Islet Cell Antibodies (ICA)
Immunoperoxidase staining in rat and human pancreas measured in JDF units.
GAD65 Autoantibodies(GAD)
Immunoprecipitation of in vitro transcribed/translated [35S-Met] labeled antigen using patient serum. [CV: inter-assay: 13.2%; intra–assay: 12.2%]
IA-2 Autoantibodies(IA2)
Immunoprecipitation of in vitro transcribed/translated [35S-Met] labeled antigen (ICA512bdc construct) using patient serum. [CV: inter-assay: 9.5%; intra–assay: 12.4%]
Insulin Autoantibodies (IAA)
Radioimmunoassay (Protein A-based) [CV: inter-assay: 19.4%; intra-assay: 8%]

* Assay statistics are taken from Pietropaolo et. al. Cytoplasmic islet-cell antibodies remain valuable in defining risk of progression to type 1 diabetes in subjects with other islet auto antibodies. *Pediatric Diabetes*, 6:184-192, (2005).

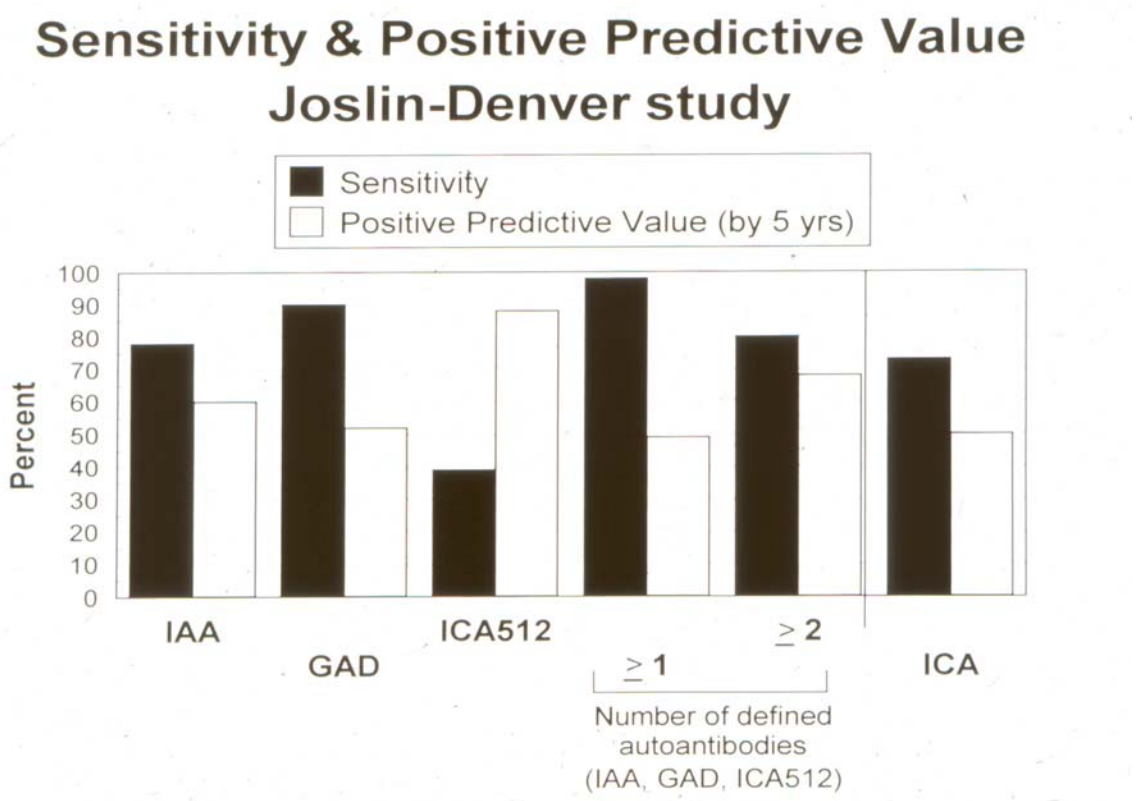

Figure 5-1 Sensitivity and positive predictive value of antibodies associated with development of Type 1 diabetes from Joslin-Denver study (Verge,et al., Diabetes 45(7):926-933,1996)

However, the great variability in the development of Type 1 diabetes makes it hard for precise prediction. All previous prognostic models for this disease used the prognostic factors recorded at one single time point for each subject to predict survival. Pietropaolo, Yu, et. al., (2005) examined a cohort of 1484 first-degree relatives (FDRs) of T1DM probands from the Children's Hospital of Pittsburgh Registry. They provide evidence that a subgroup of ICAs predicts a more rapid progression to insulin-requiring diabetes in GAD65 and IA-2 AA positive relatives and should remain part of the assessment of T1DM risk for intervention trials. In Type1 diabetes, the clinical situation usually changes with time. Figure 5.2 shows how ICA levels for the same subjects changes over time during follow-up.

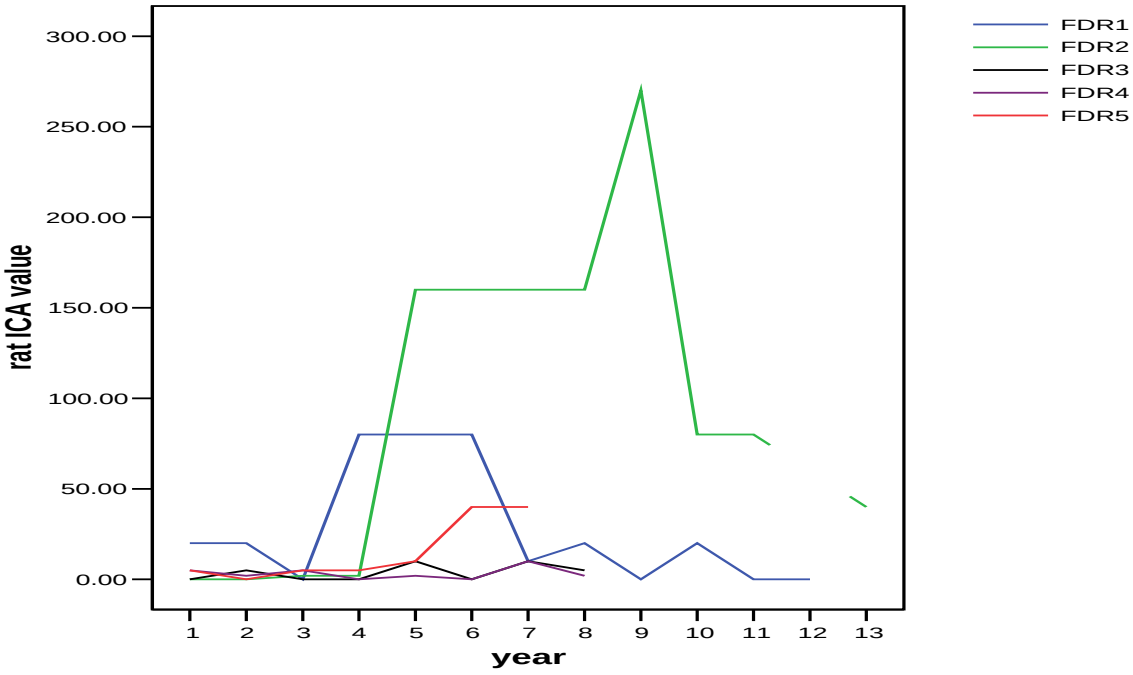

Figure 5-2 Rat ICA values over follow-up time for 5 randomly chosen subjects from a cohort of first degree relatives of Type 1 diabetic probands

The motivation of the analysis presented in this dissertation derives from existing research questions posed in the examination of the every two year dataset. This dataset consists of 499 first-degree relatives of Type 1 diabetes patients recruited from August, 22, 1977 to

August 23, 2001 at the Children's Hospital of Pittsburgh, who had at least one of the three autoantibodies (ICA, GAD, and IA2) positive for at least one time during follow-up. Figure 5.3 shows from where these subjects were selected. From these relatives, a sample of 300 were chosen using the criteria that all the baseline covariate values available, and at least one of the ICA values at year 2 or year 4 available. For each of the subjects, both demographic and clinical characteristics were recorded when they got enrolled into the study, including age, race, gender, relation to the index case, and autoantibody levels, The demographic and clinical characteristics of the study population are presented in Table 5.2 and 5.3. Attempts were made to contact the relatives by phone or mail annually to decide their diabetes status, and to obtain blood samples at approximately two year intervals to provide follow-up measurements on the antibodies.

In our previous study, we could only predict survival from baseline measurements that do not take into account changing values of the autoantibodies with time. For example, we could only measure differences in the risk of developing diabetes between groups described by the levels of the autoantibodies at the time of first blood draw. Since this new dataset recorded the antibody values (RAT ICA) for different time points in the follow-up, it gives us the ability to directly estimate the change of the risk by the change of the antibody levels for a given subject. This information will allow dynamic adjustment of risk according to a subject's antibody status, so as to more accurately identify a subset of relatives with sufficiently high-risk for Type 1 diabetes to begin preventive trials, and to identify those autoantibody-positive relatives who are unlikely to progress to diabetes. This dataset was analyzed using the analytical approaches described in the earlier sections.

We first imputed the missing covariate values and then performed the tree-structured regression method. The variables employed in the model are shown in Table 5.4. The results are

shown in Figures 5.4 and 5.5. It can be seen that, when using GAD, IA2 and ICA as predictors, the best way to separate the study population into subgroups with distinct survival time is to use GAD and ICA values. We provide evidence that when both GAD and IA2 measurements are available; ICA still appears to be an important predictor of Type 1 diabetes. As can be seen from Figure 5.5, the group of FDRs with baseline GAD greater or equal to 0.1 and baseline ICA great or equal to 120 JDF units has the greatest risk of developing T1DM in eight years. These results confirm the earlier result from a cross-sectional study where a cohort of 1484 first-degree relatives (FDRs) of T1DM probands from the same registry was analyzed (Pietropaolo, Yu, et. al, 2005).

Table 5-2 Demographic characteristics of the study population at enrollment (N=300)

Variable	Categories	Percent
Age	0-12	26.5
	12-18	11.6
	Above 18	61.9
Race	White	94.4
	Black	5.0
	Other	0.6
Gender	Male	43.7
	Female	56.3

Table 5-3 Clinical characteristics of the study population at enrollment (N=300)

Variable	Percent Positive
Rat ICA	56.2
GAD	28.9
IA2	12.6
IAA	14.5

Table 5-4 Variables employed in the tree model

Gender	
Race	
Age	Age at year 0
GAD	GAD65 Autoantibodies(GAD) year 0
IA2	IA-2 Autoantibodies(IA2) level at year 0
RAT ICA*	Islet Cell Antibodies (RAT ICA) at year 0, 2, 4

* At least one available at year 2 or 4.

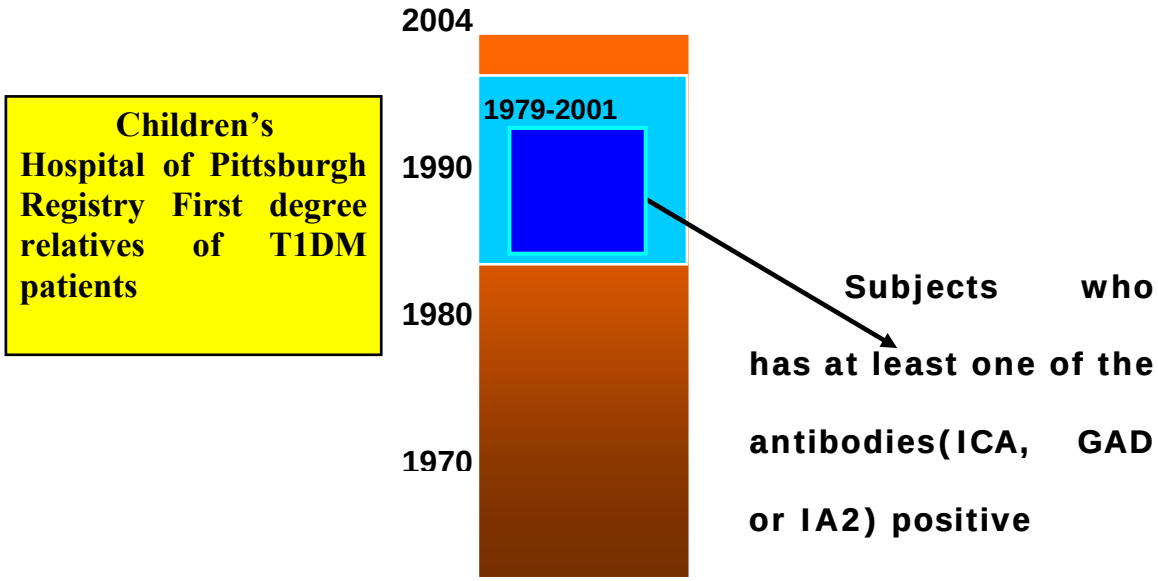

Figure 5-3 A chart of the Pittsburgh registry and the study population

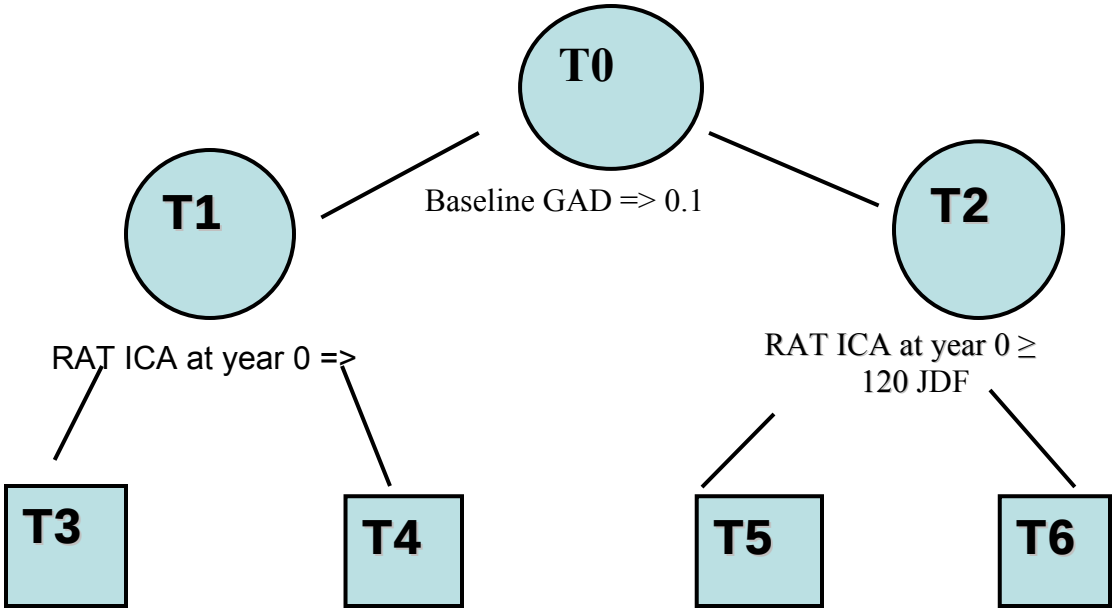

Figure 5-4 Survival tree grown using baseline and time-dependent autoantibody with log rank statistic as the splitting criterion. The split value is given below each node.

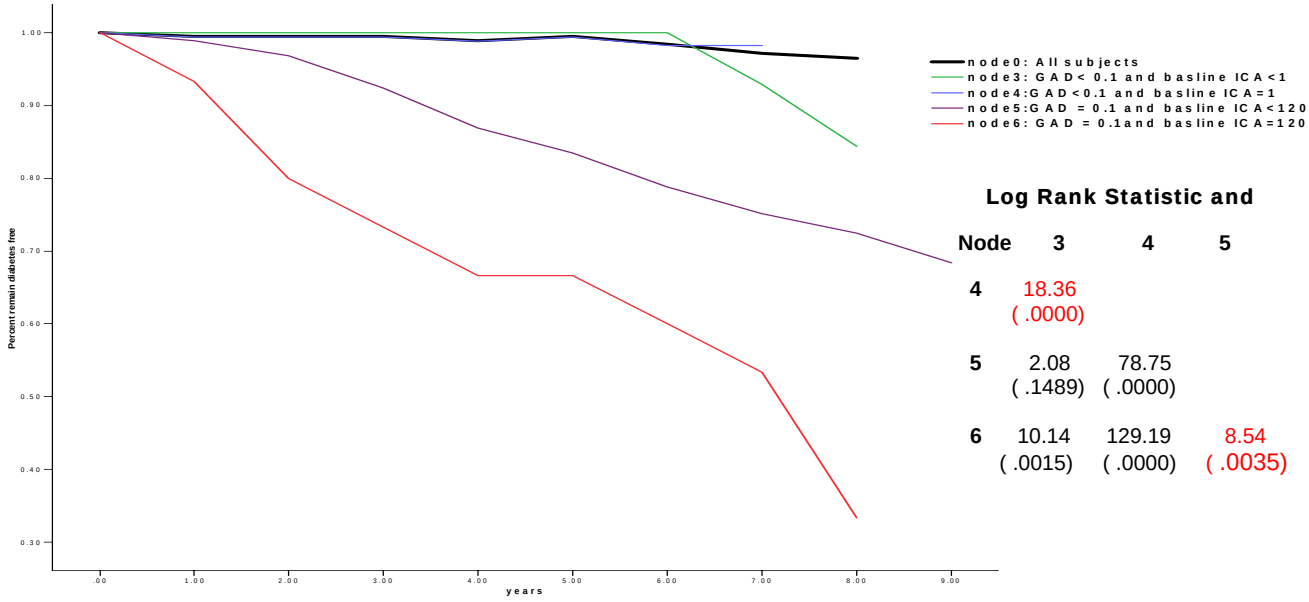

Figure 5-5 Survival curves for all subjects and subjects in the terminal nodes.

6.0 CROSS-VALIDATION STUDY

We estimate the prediction accuracy of the decision tree survival model by measures of sensitivity, specificity, and accuracy. Sensitivity is the conditional probability of correctly predicting a subject as a converter to diabetes given he/she is truly a converter. Specificity is the conditional probability of correctly predicting a subject as a nonconverter given he/she did not develop the disease. Accuracy is the probability of correctly predicting a subject's disease status. We estimated these quantities by creating a training/test partition of our dataset. In this case, we derive training/test partition of our dataset on a 2:1 ratio, many times, with replacement, a process terms bootstrapping. With each bootstrap training set we fit a tree model and test it against the corresponding test set. This validation process not only allows us to observe the empirical distributions of the prediction estimates, but also allows estimation of their variability.

After we get the bootstrap datasets, the training datasets were used to set up the survival tree model, and the cases in the corresponding test datasets were used to get the prediction measures. We define those people who develop Type 1 diabetes within eight years as positive cases and those who do not as negative cases. Since the actual disease statuses and times are already known, the predicted values can be compared with the true values, i.e., a classification table can be calculated and the prediction measures can be calculated as shown in Table 6.1. The computer program developed by Arena et al. (2004) is used in the study.

Table 6-1Calculation of prediction measures*

	Patients with disease	**Patients without disease**
Test is positive	**a**	**b**
Test is negative	**c**	**d**

***sensitivity = a / (a+c)**
specificity = d / (b+d)
accuracy= a+d/a+b+c+d

Table 6-2 Summary of prediction measures for the survival tree model*

	Minimum	Maximum	Mean	Std. Deviation
Sensitivity	.00	.71	.34	.15
Specificity	.82	.98	.90	.04
Accuracy	.76	.89	.83	.04

* For the survival tree model using Figure 5.4 and 8-year survival experience.

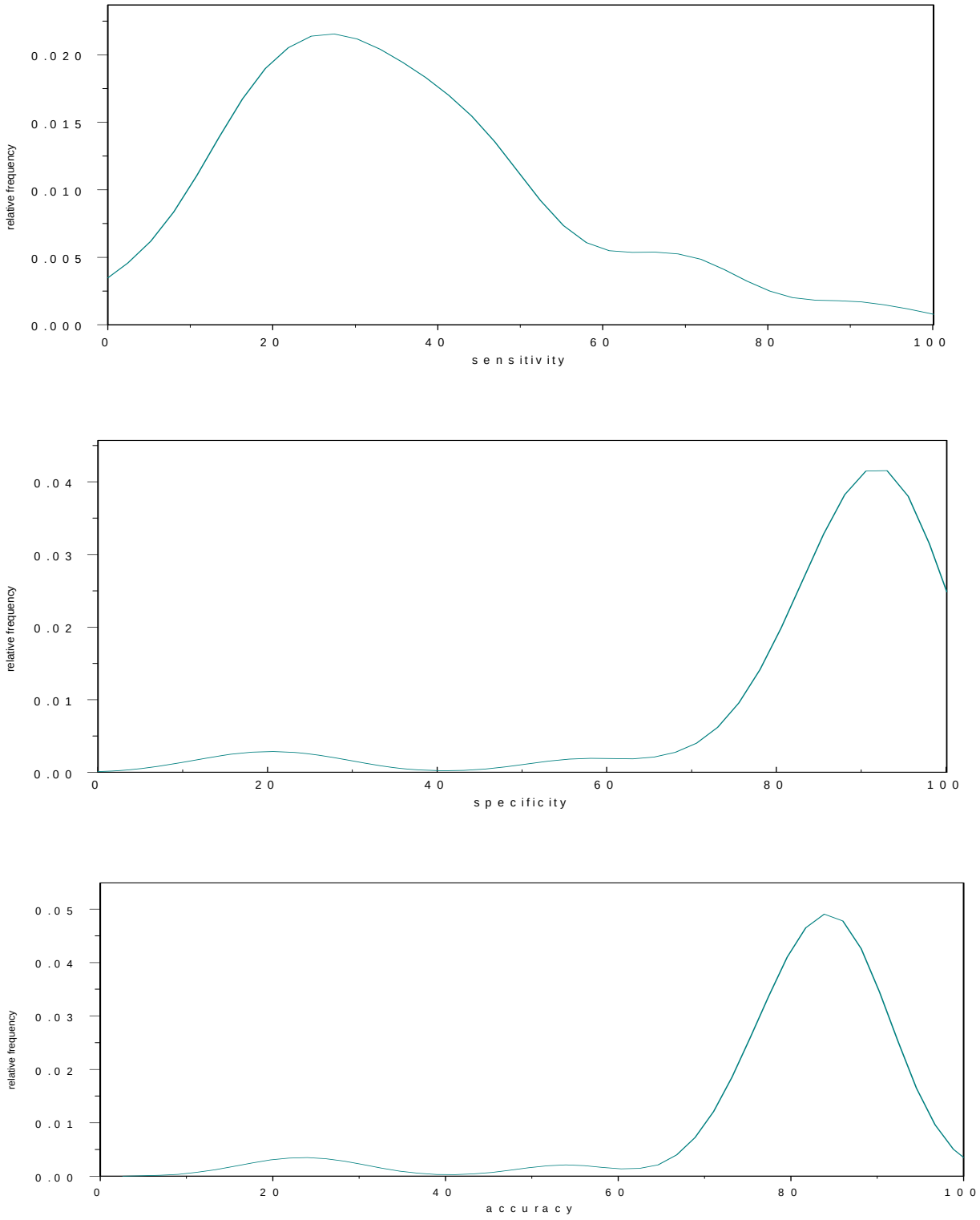

Figure 6-1 Empirical distributions of prediction measures for the survival tree model

The results are presented in Table 6.2 and Figure 6.1. We see that the distributions are skewed: sensitivity toward low values and specificity toward much higher values .We note here that the dataset is relatively small and has more negative cases than positive ones, and about 80% of the cases are censored. The missing data imputation procedure added additional variation to the data. Based on these conditions, it was difficult to uncover a strong prediction signal. Future models with higher number of cases and a more balanced ratio of positive and negative cases are expected to be more likely to express a more accurate mechanism. The models will then be medically more focused and more likely to achieve accurate predictions. In conclusion, we were able to derive modestly predictive models with a pretty high specificity and accuracy using the novel approach adapted in this dissertation.

7.0 DISCUSSION

A tree-structured survival model with incomplete and time-dependent covariates proposed here to analyze Type 1 diabetes data demonstrated moderate prediction capabilities in the cross-validation study. When we have the information about the distribution of survival time, a parametric method is clearly the best way to analyze the data. However, when the distribution is assumed incorrectly, the parametric model is not appropriate and performs poorly. However, the tree structured survival model approach has a number of important benefits. It is free from restrictive classical assumptions and does not assume any distributions for the survival data. In addition, the extraction of clinically meaningful strata, as provided by a tree-structured model, is an endpoint commonly sought by medical investigators. The tree structure also provides ready interpretability and easy classification of new patients. The method can reveal observation–based cutoff values for continuous variables. In the present example, a cutoff value for GAD65 autoantibody has been identified as 0.1. Interactions are readily recognized and no problems arise in dealing with variables of continuous or discrete type.

However, like any other methods, there are limitations to this method, and should be used as an alternative when the assumptions necessary for conventional methods cannot be made. One of the fundamental assumptions is that death time and censoring time are conditionally independent. Given the prognostic explanatory variables X's, this assumption may not be always true. Another drawback is that the tree method requires large data sets to perform well

and is computationally intensive; however, the sample size needed for the algorithm to operate optimally has not been established. This issue should be investigated with the help of simulation studies. Comparisons with other parametric approaches also depend on simulation studies. In the present approach, only univariate splits are allowed, no linear combinations of covariates are allowed. i.e. each split depends on the value of only a single predictor variable. There might be better splits based on more than one variable. We believe that the method adopted in this thesis can be extended to include splits based on linear combinations of the covariates. Methodologies for these extensions should be developed in future research.

We conclude that the tree-structured model provide a novel analytical approach to this area of research. We find that a subgroup of ICAs predicts a more rapid progression to insulin-requiring diabetes in GAD65, and ICA positive relatives. These results confirm the earlier result from a cross-sectional study where a cohort of 1484 first-degree relatives (FDRs) of T1DM probands from the same registry was analyzed (Pietropaolo, Yu, et. al, Pediatric Diabetes 2005:00: 1-9) and provide more evidence that ICAs should remain part of the assessment of T1DM risk for intervention trials.

BIBLIOGRAPHY

Ahn H and Loh WY. Tree-structured proportional hazards regression modeling. *Biometrics*, 50:471-484, (1994).

Arena VC, Sussman NB, Mazumdar S, Yu S and Macina OT. The utility of structure-activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision tree models. *SAR and QSAR in Environmental Research*, 15(1):1-18, (2004).

Bacchetti P and Segal MR. Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDs. *Lifetime Data Analysis*, 1:35-47, (1995).

Breiman L, Friedman JH, Olshen RA and Stone, CJ. Classification and regression trees. *Chapman & Halls/CRC Press*, (1998).

Conversano C, Siciliano R. Tree based classifiers for conditional incremental missing data imputation. Technical Report, Department of Mathematics and Statistics, University of Naples. Naples, Italy, (2003).

Cox DR. Regression models and life tables with discussions. *J. Royal Stat. Soc. Ser. B*, 34:187-220, (1972).

Davis RB and Anderson JR. Exponential survival trees. *Statistics in Medicine*, 8:947-961, (1989).

Ford I, Norrie J and Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine*, 14:735-746, (1995).

Gail MH, Wieand S and Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431–444, (1984).

Gehan EA. A generalized wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, 52:203-223, (1965).

Gordon L. and Olshen RA. Tree-structured survival analysis. *Cancer Treatment Reports*, 69:1065-1069, (1985).

Harrington DP and Fleming TR. A class of rank test procedures for censored survival data. *Biometrika*, 69:553-566, (1982).

Kaplan EL and Meier P. Nonparametric estimation for incomplete observations. *J. Am. Stat. Assoc*., 53:457-81, (1958).

Lancaster T and Nickell S. The analysis of re-employment probabilities for the unemployed. *J.R. Statist. Soc. A.*, 143:141-165, (1980).

LeBlanc M and Crowley J. Relative risk trees for censored survival data. *Biometrics*, 48:411-425, (1992).

LeBlanc M and Crowley J. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457-467, (1993).

Morgan JN and Sonquist JA. Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association*, 58:415-434, (1963).

Peto R and Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*, Series A (General), 135(2):185-207, (1972).

Pietropaolo M, Yu S, Libman IM, Pietropaolo SL, Riley K, LaPorte RE, Drash AL, Mazumdar A, Trucco M and Becker DJ. Cytoplasmic islet-cell antibodies remain valuable in defining risk of progression to type 1 diabetes in subjects with other islet auto antibodies. *Pediatric Diabetes*, 6:184-192, (2005).

Prentice RL. Linear rank tests with right censored data. *Biometrika*, 65(1):167-169, (1978).

Segal MR. Regression trees for censored data. *Biometrics*, 44:35-47, (1988).

Struthers CA and Kalbfleisch JD. Misspecified proportional hazards models. *Biometrika*, 73:363-369, (1986)

Tarone RE and Ware J. On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1):156-160, (1977).

Therneau TM, Grambsch PM and Fleming TR. Martingale-based residuals for survival models. *Biometrika*, 77:147-160, (1990).

Verge CF, Gianani R, Kawasaki E, Yu L, Pietropaolo M, Jackson RA, Chase PH, Eisenbarth GS. Prediction of Type 1 diabetes in first-degree relatives using a combination of insulin, GAD and ICA512bdc/IA-2 auto antibodies. *Diabetes*, 45(7):926-933, (1996).